Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Unveiling Probability Histograms from Random Signals using a Variable-Order Quadrature Method of Moments

Menwer Attarakih[a,*], Mark W. Hlawitschka[b], Linda Al-Hmoud[a],and Hans-Jörg Bart[c]

[a] The University of Jordan, Department of Chemical Engineering, Amman, Jordan
[b] Institute of Process Engineering, Johannes Kepler University (JKU), Linz, Austria
[c] Mechanical & Process Engineering, RPTU Kaiserslautern, Kaiserslautern, Germany
* Corresponding Author: m.attarakihl@ju.edu.jo.

ABSTRACT

Random signals are crucial in chemical and process engineering, where industrial plants generate big data that can be used for process understanding and decision-making. This makes it necessary to unveil the underlying probability histograms from these signals with a finite number of bins. However, the search for the optimal number of bins is still based on empirical optimisation and general rules of thumb. In this work, we introduce an alternative and general method to unveil probability histograms. Our method employs a novel variable-order QMOM, which adapts automatically based on the relevance of the information contained in the random data. The number of bins used to recover the underlying histogram is found to be proportional to the information entropy, where a search algorithm is developed that generates bins and assigns probabilities to them. The algorithm terminates when no more significant information is available for assignment to the newly created nodes, up to a user-defined threshold. In conclusion, our method is a universal histogram reconstruction technique that only requires enough numbers of moments to work. The method has been validated intensively using synthetic random signals and real-life problems.

Keywords: Probability histogram, Modelling, Simulation, Random signals, VOQMOM, Population Balances.

INTRODUCTION

Random signals are vital in chemical and process engineering, data science, and machine learning. Industrial plants collect and analyze large volumes of data for process understanding, data compression, and decision-making, modelling industry 4.0 and 5.0 transformation and equipment or even process digital twins [1, 2, 3]. However, this data is not free of instrumental errors, where signal processing corrects it, enhances data accuracy and reliability. Despite advanced techniques that have been available, basic methods are commonly used due to resources' limitations which still depend on basic statistical representation methods [1]. In the sense of data compression, probability histograms are widely used as non-parametric density estimators, essential for visualizing data and deriving summary quantities like entropy from the underlying density, determining mechanical and physiochemical properties of particles from industrial systems and for real time online monitoring and control purposes [1]. However, the accuracy of these quantities is influenced by the number of bins selected for the histogram, which, in turn, determines the bin width based on the data range. Moreover, the empirical assumption of using uniform bin width has no solid theoretical support, where in most cases it is used for the simplicity of mathematical treatment [4, 5]. Unfortunately, finding the optimal number of bins is still based on empirical optimization and general rules of thumb. These methods are based on empirical formulas for the uniform optimal bin width by minimizing the integrated mean squared error of the histogram model H(x) as compared to the true and unknown underlying density f(x). Two popular formulas after Freedman and Diaconis [5] and that of Scott [6] were commonly used in which the bin width is inversely proportional to the cubic root of the number of data points in the set, while the proportionality constants are different but depend on the variance and the interquartile range of the data. On the other hand, one of the recent approaches is centered around a data-based method to estimate the optimal uniform bin-width of histograms using an optimization algorithm but not an

explicit empirical formula [4]. The disadvantages include the potentially large number of bins, the uniform bin distribution regardless of weights, and the inconsistency of the histogram with the low-order moments of the true distribution.

In this contribution we took a different approach that is free of the two main drawbacks in the previous work: The histogram's bin width and number of bins. We relaxed the assumption of uniform bin width and the number of the bins are evolved based on the available feasible information in the data set according to Shannon information entropy: $S = -\sum[p_i \ln(p_i)]$ with the constraints that $m_0 = \sum[p_i] = 1$ where m_0 is the normalized zeroth moment of the data set and p_i is the i^{th} probability associated with the i^{th} bin of the histogram. Our simple search algorithm for N (Number of data points in the data set) is constrained by the feasible range of the data $R = [x_{min}, x_{max}]$ and the relevance of the probabilities p_0 and p_{N-1} of the data in the first and last histogram bins.

THE VARIABLE ORDER QMOM

As a data compression method, the Quadrature Method Of Moments (QMOM) works based on the calculated moments of the unknown probability density function. In its original formulation, the QMOM algorithm provides closed integrals (and hence moments as special case) of the unknown pdf based on the availability of its low-order moments. It is mainly used to close integrals when the weight function is not known explicitly. Since its introduction by McGraw [7], the QMOM has been the subject of considerable theoretical and applied research during the last two decades. These applications cover wide areas which include aerosols microphysics, nanoparticle formation, turbulent mixing, dispersed phase flows, crystal morphology, cell growth and differentiation and uncertainty propagation through dynamical systems, to name but a few [8]. As an ill-conditioned inverse moment problem, there is no simple and general inversion algorithm to reconstruct the unknown probability histogram from which the moments are calculated or measured. Because of this, the QMOM is rarely used or thought of as a method that can recover the underlying histogram from the available information in the nodes and weights provided by the method itself.

The QMOM

The QMOM can be mathematically presented as an adaptive data compression method where information about the random signal is converted into a finite set of low-order moments irrespective of the size of the data set. Therefore, given a set of low-order moments $m_r, \mathrm{r} = 0, 1, \ldots 2N_q - 1$ as calculated from a pdf, the QMOM algorithm calculates a set of N_q nodes and weights (ζ, w) with the same principle as that of the Gauss-Legendre quadrature. The domain of the node set $\zeta \in (0,\infty)$ or $\zeta \in [x_{min}, x_{max}]$ depending on the bounds of the random data from which the set m_r is calculated. The QMOM algorithm used in this work is chosen based on the efficient derivation of Upadhyay [9] using the Chebyshev orthogonal polynomials. This can handle large number of nodes that are not necessarily positive when compared to the classical product-difference algorithm (PDA) that was presented by [7]. Actually, Upadhyay [9] showed that the Chebyshev-QMOM (ChebQMOM) is superior to the PDA and is found to be more robust and can be used for a wider class of problems when high number of nodes is required as in our present work.

The Variable-Order QMOM Algorithm

To proceed in the derivation of the Variable-Order QMOM (VOQMOM), we have to show that the weights (w_i, i = 0, 2, ..N_q-1) as calculated by the ChebQMOM are nothing more than probabilities of the function f(x;:) that are placed at the corresponding nodal points (ζ_i, i = 1,2, ..N_q-1).

$$w_i = \int_{\zeta_{i-1/2}}^{\zeta_{i+1/2}} \mathrm{f(x;:)}dx \tag{1}$$

where ($\zeta_{i-1/2}$, $\zeta_{i+1/2}$) are unknown grid boundaries, which will be referred to as histogram bins. The bins boundaries should satisfy the inequality ($\zeta_{i-1/2} < \zeta_I < \zeta_{i+1/2}$) and the bins are not necessarily equal in width. The estimation of these grid boundaries is given by: $\zeta_{i+1/2} = \frac{1}{2}(\zeta_{i-1} + \zeta_i)$ which are functions only of the low-order moments of the given random signal. To reconstruct the probability histogram $H(\zeta_i)$ with probability (w_i) which is placed at (ζ_I), we make use of the integral of the square of errors as defined by [4]:

$$E(H, f) = \int_0^\infty [H(x) - f(x)]^2 dx \tag{2}$$

Since the nodal points (ζ_I) are roots of orthogonal polynomials by definition of ChebQMOM, we can close the integral of Eq.(2) as follows:

$$E(H, f) = \sum_{j=0}^{N_q} \left(w_j f(\zeta_j) - w_j H(\zeta_j) \right) \tag{3}$$

In Eq.(3), we have forced the histogram probability $H(x)$ to have low-order moments that are equal to those of $f(x)$, which is an enough assumption to deem the consistency of the reconstructed histogram w.r.t. the underlying density from which the moments are calculated. Now, for the error in Eq.(3) to vanish at the collocation points (ζ_I), it is sufficient to have:

$$H(\zeta_i) = f(\zeta_i) = \frac{1}{\Delta\zeta_i} \frac{w_i}{\sum w_i} \tag{4}$$

where $\Delta\zeta_i = \zeta_{i+1/2} - \zeta_{i-1/2} = \frac{1}{2}(\zeta_{i+1} - \zeta_{i-1})$.

It is not difficult to show that the order of approximation of $H(\zeta_i)$ in Eq.(4) is $O(\Delta\zeta_i^2)$ because the approximation in Eq.(4) is equivalent to using the midpoint rule to

approximate the integral in Eq,(1). Note that Eq.(4) provides a complete theoretical inversion of the unknown probability density function f(ζ_i,:). The order of convergence is proportional to the square of the distance between the (i-
1) and (i+1) bin centers divided by 4.

The question now is how many bins (or quadrature nodes) that are sufficient to resolve the histogram probability $H(x)$ given a set of low-order moments that are computed from the random signal? The answer is nontrivial as it is related to the Shannon information entropy based on the available probabilities (w_i):

$$S = -\sum_i w_i \ln(w_i) \quad (5)$$

From Eq.(5), it is evident that as the number of bins used to recover the underlying histogram increases, so does the information entropy, until no more information is required. This implies that adding extra bins will not contribute any relevant information to the histogram, as indicated by very low probability values, signifying the irrelevance of newly added nodes. In the hypothetical limit where the data has zero information entropy (Dirac delta function), the number of bins is reduced to one. In the variable order QMOM realm, the number of bins is determined through a simple search algorithm that optimally assigns nodes through sampling the unknown function or process from which the random data is generated. The algorithm terminates when no more significant information is available for assignment to the newly created nodes, up to a user-defined threshold. If the data originates from a dynamic source with varying mean and variance, the number of bins and its boundaries will adjust itself to reflect the nature of the dynamic data. The complete detailed algorithm is shown in Figure 1. Note that the "while loop" implements the simple search and variable order of the ChebQMOM algorithm where the order is determined if the last computed weight (probability) "W_{N-1}" is less than very small, specified tolerance (Tol) and if the maximum and minimum computed nodes are outside their physical range. The algorithm is initiated by providing a vector of computed low-order moments "MOM", initial order "N = 2", Tol, x_{min} and x_{max}. The output is a set of computed nodes and weights (ζb, wb). These conditions are found in all simulated cases to correspond to the maximum entropy, as given by Eq.(5). After reaching this point, the entropy "S" begins to decrease because the weights at least on one end of the distribution fall to zero. In Figure 4, the VOQMOM process is illustrated with a real-life sample showing the workflow and the significance of the random signal identification with 22 low-order moments associated with the reconstructed histogram. Also, the plateau of the information entropy is clear near the maximum number of bins. Note that the ChebQMOM is implemented in an efficient way to reduce the numerical round-off errors using MATLAB.

RESULTS AND DISCUSSION

In this section we will present sample of results using synthetic random signals and real-life case studies. Also, in addition to the static sampling, we present a dynamic case study by solving the simultaneous particle aggregation and growth in a closed batch process.

```
" Input MOM vector of low-order moments, Tol = small "
" ζmin = minimum random variable size "
" ζmax = maximum random variable size "
N ← 2
M_max ← max(MOM)
for r ∈ 0..2·N − 1
  M_r ← MOM_r
(ζ  W) ← ChebQMOM(M,N)
while (M_max·W_0 ∧ M_max·W_N−1) ≥ Tol ∧ N < N_max ∧ (ζ_0 ≥ 0 ∧ ζ_N−1 ≤ ζmax)
  N ← N + 1
  for r ∈ 0..2·N − 1
    M_r ← MOM_r
  (ζt  Wt) ← ChebQMOM(M,N)
N ← N − 1  if  ζ_0 < ζmin ∨ ζ_N−1 ≥ ζmax
for r ∈ 0..2·N − 1
  Mp_r ← MOM_r
(ζb  Wb) ← WNODES(Mp,N)
```

Figure 1. Variable-order ChebQMOM algorithm.

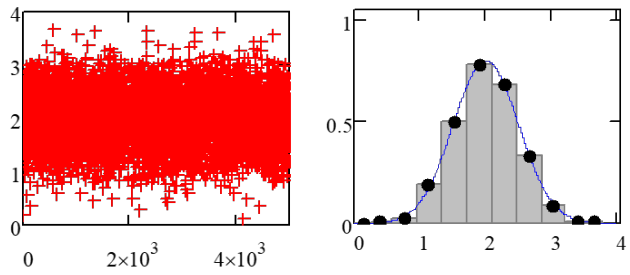

Figure 2. Unveiling random signal produced from normal distribution (mean = 2, std = 0.5) using 5000 samples. On the right-hand side is the recoverd probability histogram using 11 bins with 22 low-order moments.

This case study demonstrates how the VOQMOM identifies the data that is coming from a dynamic source with varying mean and variance where the boundaries of the bins move in a dynamic way. In addition to this, similar case studies were conducted on particle breakage in closed and open systems, particle breakage and coagulation, an industrial scale mixer where a light liquid phase is dispersed in water as the continuous phase.

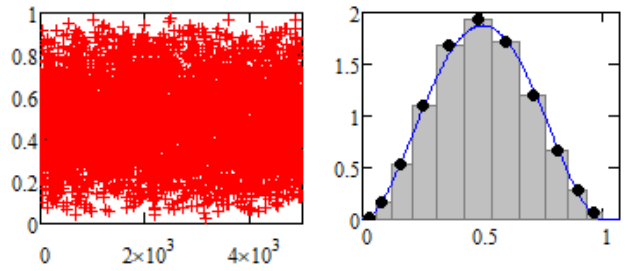

Figure 3. Unveiling random signal genrated from a beta distribution (parameters a = 3 and b = 3) using 5000 samples. On the right-hand side is the recovered probability histogram using 11 bins based on 22 low-order moments.

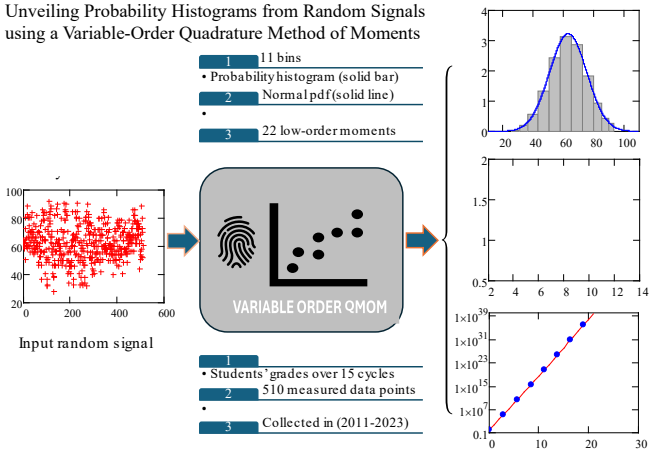

Figure 4. Moment-consistent histogram unveiled from students' grades using the Variable-Order QMOM (VOQMOM). To the left is the random signal of students' grades during 15 cycles. In the center is the VOQMOM process, and on the top right is the unveiled density histogram (with 11 bins) compared to the normal pdf with the same mean and variance as the data set. On the right (middle) is the information entropy of the data set as function of number of bins, while on the right bottom is the predicted (filled circles) and original set of moments (solid line).

Unveiling probability histograms from Synthetic random signals

A sample of three case studies that were synthesized to prove the concept of the VOQMOM using random sampling from normal, beta and lognormal probability density functions (pdf). The random variable "x" is normalized using the relation: $u = \frac{x - x_{max}}{x_{max} - x_{min}}$ and the moments of the sampled data were calculated using the relation $m_r = \frac{1}{N}\sum_{j=0}^{NoSamles} u_j^r \ , r = 0, 1, \ldots \ NoSamples$. This moment transformation results in a moment space that is decreasing exponentially showcasing the irrelevant information contained in the high-order moments which are usually address the information content in the tail of the pdf. Figure 2 (Left) shows the 5000 random samples that are drawn from a normal probability density function with mean = 2, and std = 0.5. The VOQMOM algorithm identified this random signal using the algorithm of Figure 1 which discovered an 11-bin probability density histogram as compared to the original normal pdf showen in Figure 2 (Right). In addition to this, the histogram is consistent w.r.t. 22 low-order moments of the original data. Similarly, Figure 3 shows the results of identifying 5000 random samples drawn from the beta probability density function by the VOQMOM ending with an 11-bin histogram. In Figure 5, the VOQMOM is challenged by 5000 random samples that are drwan from the unsymmetrical log pdf (mean = 1.349 and std = 0.419). Again, with 11 discovered bins, the unveailed probability density histogram is very close to the orignal pdf.

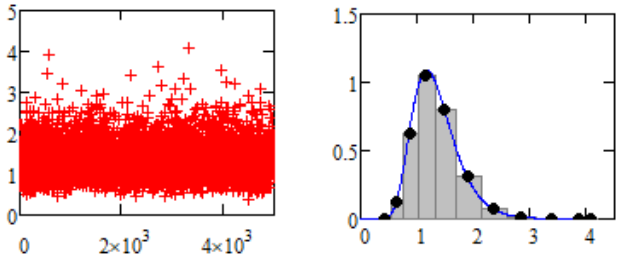

Figure 5. Unveiling random signal genrated from a lognormal probability density function (mean = 1.349 and std = 0.419) using 5000 samples. On the right-hand side is the recovered probability histogram using 11 bins based on 22 low-order moments.

Table 1: Unveiled probability histograms using the VOQMOM based on different real-life random data.

Case	Description	No of bins
1	mean maximum air to surface temperature in Amman	13
2	adult diastolic blood pressure	11
3	abalone total weight data	12
4	students' grades for 15 cycles.	11
5	mean wind speed data	12
6	droplet size distribution data	11

From these results, it is evident that the VOQMOM discovers incrementally new nodes and decides to stop when the individual probability carried by the terminal nodes (first & last nodes) becomes less than a user specified tolerance (Tol = 10-12). Also, the physical bounds placed on the terminal nodes is checked accordinglly where the probabilities of the terminal nodes approach zero when the physical bounds are exceeded.

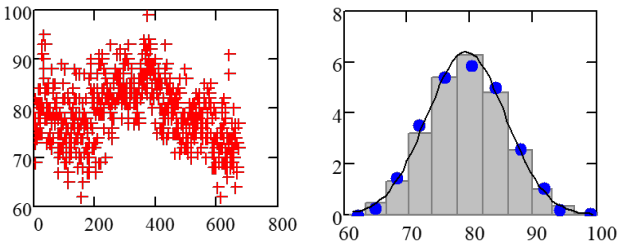

Figure 6. Unveiling random signal based on measured diastolic blood pressure of a male adult (age 50-56) using 671 samples. On the right-hand side is the recoverd probability histogram using 11 bins based on 22 low-order moments as compared to the normal pdf.

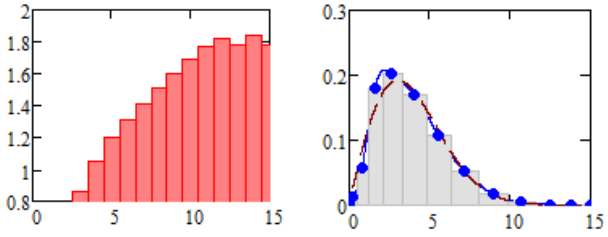

Figure 7. Unveiling random signal based on measured mean wind speed [10] using ~35000 samples. On the left-hand side is the information entropy as function of bins number and on the right-hand side is the recoverd probability histogram using 12 bins based on 24 low-order moments., The dashed line is the Weibull pdf with shape and scaling parameters calculated based on the correlations from [10].

Unveiling probability histograms from real-life random signals

Having validated the concept of the simple search algorithm of the VOQMOM using random signals from a reference pdf, we demonstrate the validity of the method using real-life data sets. Figure 6 (Left) shows 671 random measurements of diastolic blood pressure for an adult male (50-56 years old) during a time span of 6 years. The search algorithm of the VOQMOM identified an 11-bin probability histogram behind these random data set as shown in Figure 6 (Right). The histogram is consistent w.r.t. 22 low order moments. Comparing this unveiled histogram to the normal pdf, it is clear that the data is well presented by this pdf with mean and standard deviation as those of the original data. In the second real case study, a random measured data set of size ~ 35000 which represents mean wind speed data [10] as shown in Figure 7. On the left-hand side of this figure is the information entropy as function of bins number. It is clear when new node or bin is added, the information entropy increases until a point where the probabiltiy of the added node approaches zero. At this point we observed that the entropy drops slightly in all the investigated case studies signifiying the termination of the algorithm at the maximum possible node number. On the right-hand side of Figure 7 is the recoverd probability histogram using 12 bins based on 24 low-order moments. Also, the results are compared to the Weibull pdf (dashed line) with shape and scaling parameters calculated based on the correlations derived by Shu and Jesson [10] using mean wind speed data. It is clear that this data set is well represented by the Weibull pdf. The last case study consists of a data set of abalone total weights in mg that is measured from 4177 individuals [12]. Figure 8 (Left) shows the increase in the information entropy until a flat plateau is reached near the best number of bins (12). The unveiled histogram from these random data is shown in Figure 8 (Right) and is compared to the Weibull distribution with shape and scaling factors that are calculated by correlations from [10]. It is noteworthy to mention here that the optimal number of uniform bins obtained from this data set was found to be 14 using an optimization algorithm of Knuth [4]. Furthermore, Table 1 shows many cases studies which are based on real-life experience with main observation that the number of bins predicted by the VOQMOM ranges from 11 to 14.

Unveiling probability histograms by solving the population balance equation

In this case study we tested the VOQMOM for the dynamic evolution of simultaneous particle growth and coagulation under constant coagulation rate and growth velocity given by $G(x) = G_0 x/3$. From the initial condition $f(x) = 3x^2 \exp(-x^3)$, the number of bins is found to be 11. After 10 seconds of simulation time using Heun's method with a time step of 0.025 seconds and Λ = 0.75 (ratio of coagulation to growth rate), the number of bins discovered by the VOQMOM is 13. Figure 9 shows the normalized histogram (w.r.t. zeroth moment) at t = 0 (Left) and at t = 10 seconds (Right). All the 22 low-order moments were calculated with high accuracy when compared to the analytical solution of Gelbard and Seinfeld [11]. In spite of the growth of the particle size around 20 times of its initial value, the solution at t = 10 s is free of numerical diffusion due to the integral representation of the growth term thanks to the QMOM. Note that the integral source term in the population balance equation is integrated with a variable order QMOM at each instant of time. The CPU time using Heun's method for this case is in the order of 5 seconds using a state of the art msi-laptop (Intel(R) Core(TM) i7, CPU at 2.6 GHz & 8 GB RAM). The dominance of coagulation over particle growth is manifested by the increase of information entropy from 1.68 to 1.77 that is reflected by the increase of nodes from 11 to 13.

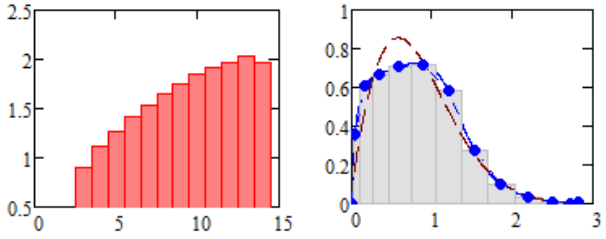

Figure 8. Unveiling random signal based on measured abolane total weight in mg using 4177 samples. On the left-hand side is the information entropy as function of bins' number and on the right-hand side is the recoverd probability histogram using 12 bins based on 24 low-order moments as compared to Weibull pdf (dashed line) with shape and scaling parameters calculated based on the correlations of Shu and Jesson [10].

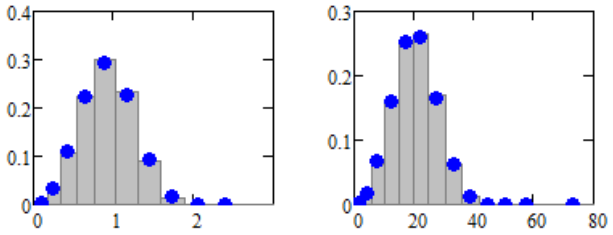

Figure 9. Particle growth & coagulation with Λ = 0.75 in a simplified batch crystallizer. Left: Normalized histogram at t = 0 with 11 bins. Right: Normalized histogram as compared to the exact ones (filled circles) with number of moving bin centres = 13 at t = 10 s.

SUMMARY AND CONCLUSIONS

The VOQMOM is a non-parametric method that can be used to identify general probability histograms by processing big random data which is drawn from real-life problems. It works through compressing a given set of random data into low-order moments, which are then processed to encode the underlying pdf from which the data is generated. It uses a simple linear search algorithm coupled with ChebQMOM and needs only a sufficient number of nodes and weights to work with negligible computational cost. Using different analytical and practical case studies, it is found that 10 to 14 bins were sufficient to reconstruct the pdf with a local accuracy of $O(\Delta\zeta_i 2)$. The algorithm terminates when no more informaiton is available to assign to the terminal nodes.

REFERENCES

1. Thibault, E., Chioua, M., McKay, M., Korbel, M., Patience, G. S. and Stuart, P. R. Experimental methods in chemical engineering: Data processing and data usage in decision-making. Can. J. Chem. Eng. 101: 6055-6078 (2023).
2. Sarramaigna, P., Besbes, M., Zolghadri, M. and Sadoul, P. O. Modelling Industry 4.0 transformation: A comparative approach between academic literature and French companies' transformation cases. Procedia CIRP 120:786-791 (2023).
3. Peterson, L., Gosea, I. V., Benner, P. and Sundmacher, K. . Digital twins in process engineering: An overview on computational and numerical methods. Computers & Chemical Engineering 193:108917 (2025).
4. Knuth, K. H. Optimal data-based binning for histograms and histogram-based probability density models. Digital Signal Processing, 95: 102581 (2019).
5. Freedman, D. and Diaconis, P. On the Histogram as a Density Estimator: L 2 Theory Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 57: 453–476 (1981).
6. Scott, D. W. Sturges' rule. Wires Computational Statistics 1: 303–306 (2009).
7. McGraw, R. (1997). Description of aerosol dynamics by the quadrature method of moments. Aerosol Sci. & Tech. 27:255-265.
8. Attarakih, M., Bart, H.-J. and Abu-Khader, M. On the solution of the population balance equation: From global to local constrained maximum entropy method. Chem. Eng. Sci. 209:115168 (2019).
9. Upadhyay, R. R. Evaluation of the use of the Chebyshev algorithm with the quadrature method of moments for simulating aerosol dynamics. Journal of Aerosol Science 44:11-23 (2012).
10. Shu, Z. R. and Jesson, M. Estimation of Weibull parameters for wind energy analysis across the UK. J. Renewable & Sustainable Energy 13: 023303 (2021).
11. Gelbard, F. and Seinfeld, J. H. Numerical solution of the dynamic equation for particulate systems. J. Comp. Phys. 28:357-375 (1978).
12. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI repository of machine learning databases, http://www.ics .uci.edu /~mlearn /MLRepository.html (1998).